# Learning Hidden Markov Models for Anomaly Detection in Time Series

**Kareth León**[1] Henry Arguello[1] Jean-Yves Tourneret[2]
Collaborators: Florian Mouret, TerraNIS

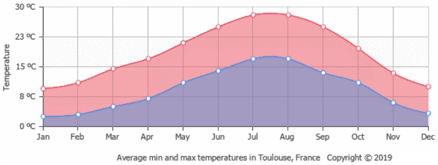
[1]Universidad Industrial de Santander, Colombia
[2]University of Toulouse, IRIT/INP-ENSEEIHT/TeSA, France

March 2020

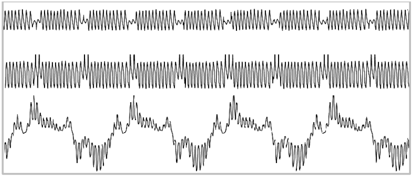
High Dimensional Signal Processing Research Group

# Context: Time Series

- A time series is a set of points indexed by time.



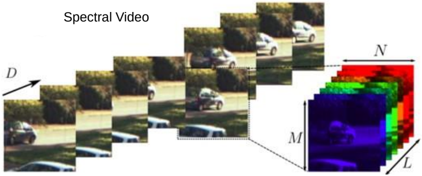

Univariate: Temperature



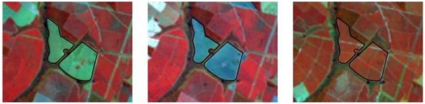
Multivariate: Telemetry time series
(temperature, pressure, voltage)

- Temporal features extracted from high-dimensional signals indexed by time (e.g. spectral video, multitemporal hyperspectral images)

Spectral Video



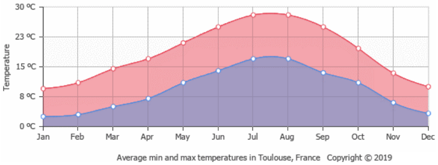

Temporal Remote Sensing



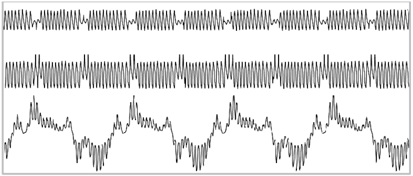
1a) Apr. 24th, 2007 1b) Jun. 11th, 2007 1c) Mar. 25th, 2008

# Context: Time Series

- A time series is a set of points indexed by time.



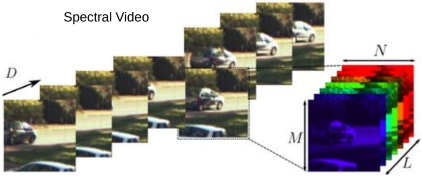

Univariate: Temperature



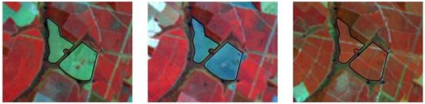
Multivariate: Telemetry time series
(temperature, pressure, voltage)

- Temporal features extracted from high-dimensional signals indexed by time (e.g. spectral video, multitemporal hyperspectral images)

Spectral Video




Temporal Remote Sensing



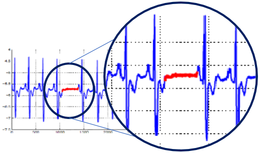
1a) Apr. 24th, 2007 1b) Jun. 11th, 2007 1c) Mar. 25th, 2008

**Goal**: Learn the temporal structure to discriminate anomalies!

# What is an anomaly?

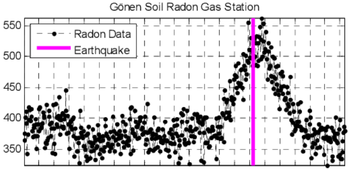
# Some Applications

- Medical: Heart rate (ECG)

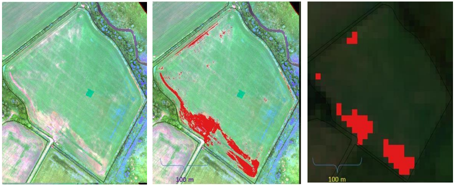


- Economy: Fraud Detection


- Seismology: Seismic Activity




- Agriculture: Crop Monitoring

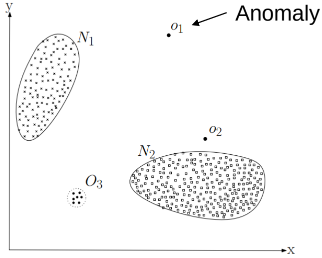


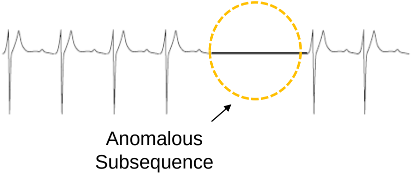

- Networks and Communication: Network Intrusion, Malware Detection, among others.

# Anomaly Detection: Type of Anomalies

### Point Anomalies

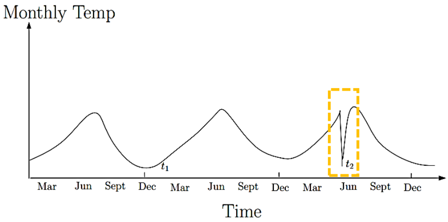
- An individual data instance is anomalous




### Contextual Anomalies

- A data instance is anomalous in a specific context (but not otherwise).

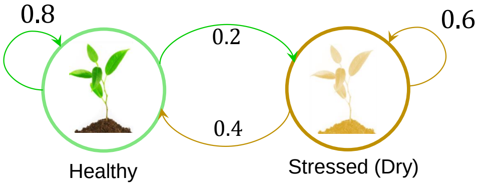



### Collective Anomalies

- A collection of related data instances is anomalous with respect to the entire data set, but not individual values (e.g. breaking rhythm in ECG)

# Hidden Markov Models (HMM): Introduction

**Markov Model (first-order)**: Stochastic model for changing systems.




**Transitions probabilities**

| | **Tomorrow** | |
|---|---|---|
| **Today** | Healthy | Stressed |
| Healthy | 0.8 | 0.2 |
| Stressed | 0.4 | 0.6 |

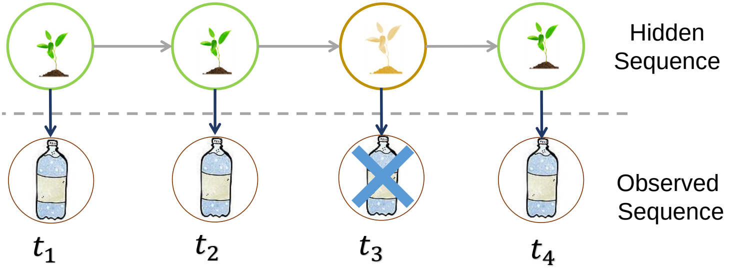
- For the sequence $\{z_1, ..., z_T\}$: $P(z_{t+1}|z_t, z_{t-1}, ..., z_1) = P(z_{t+1}|z_t)$ (Rule).
- States $S = \{\text{healthy}, \text{stressed}\}$.
- Ex: Given that today the plant is healthy $(z_t = \text{healthy})$, the probability that it can be stressed tomorrow $z_{t+1}$ is:
  $P(z_{\text{today}}|z_{\text{tomorrow}}) = P(z_{\text{today}} = \text{healthy}|z_{\text{tomorrow}} = \text{stressed}) = 0.2.$

**Hypothetical situation**

Suppose that you are allergic to plants so you cannot check the state of the plant. The only way to know the **state** of the plant is by observing if the caretaker of the garden lets a **bottle** of water or not in the kitchen.

# HMM Components

Thus, instead of observing the real state, you observe the bottle. The real state of the plant is hidden!




**Components**

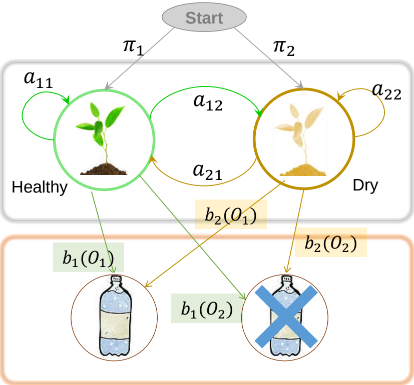
1. States: $S = \{\text{healthy}, \text{stressed}\}$
2. Observations: $x = x_1, x_2, ..., x_T$
3. Initial probabilities: $\pi$
4. Transitions probabilities: $\mathbf{A}$
5. Emission probabilities: $\mathbf{B}$

**Emission probabilities**

| **Probability of observation** |  |  |
|---|---|---|
| Stressed | 0.1 | 0.9 |
| Healthy | 0.8 | 0.2 |

# Hidden Markov Models

- Observed time series: $x = [x_1, ..., x_T]$,
- Set of possible states: $S = \{s_1, ..., s_D\}$,
- Hidden sequence: $z = [z_1, ..., z_T]$,
- Number of states: $D$.




A HMM model is given by $\theta = \{\pi, \mathbf{A}, \mathbf{B}\}$.

**Initial Probabilities: $\pi$**

$$\pi_i = P(z_1 = s_i), i = 1, ..., D$$

**Transition Probabilities: $\mathbf{A}$**

Probability to go from state $i$ to state $j$:

$$a_{i,j} = P(z_{t+1} = s_i | z_t = s_j)$$

where $\{s_i, s_j\} \in S$, and $i, j \in \{1, ..., D\}$.

**Emission Probabilities: $\mathbf{A}$**

Observation probability distribution in state $i$ such that $\mathbf{B} = \{b_i(\cdot)\}$. The probability of the observations can be:

- Discrete
- Continuous

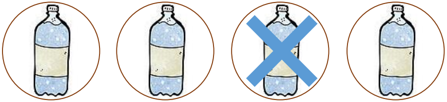
# Observation Probability Distribution $B = \{b_i(\cdot)\}$

▶ **Discrete Observations:** Observations can belong to a codebook $V = \{v_1, ..., v_K\}$, e.g., for the bottle observation Y (yes) or N (not), the codebook is $V = \{Y, N\}$.

The probability is defined as:

$$b_i(x_t) = P(x_t = v_k | z_t = s_i), \quad (1)$$

where $1 \leq i \leq D$, $t = 1, ..., T$.



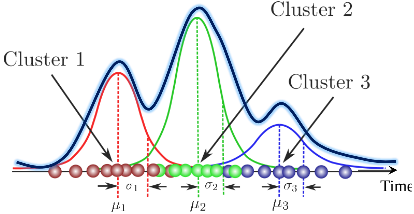
$\{Y, Y, N, Y\}$

▶ **Continuous Distribution**: Observations follow a specific distribution, e.g., a Gaussian distribution or the mixture of multiple Gaussians.

The emission probabilities are defined as:

$$b_i(x_t) = \sum_{m=1}^{M} C_{i,m} \mathcal{N}(x_t | \mu_{i,m}, \Sigma_{i,m}), \quad (2)$$

where $1 \leq i \leq D$, $M$ is the number of Gaussians, $\mu$ is the mean, $\Sigma$ is the covariance, and $\sum_{m=1}^{M} C_{i,m} = 1$.



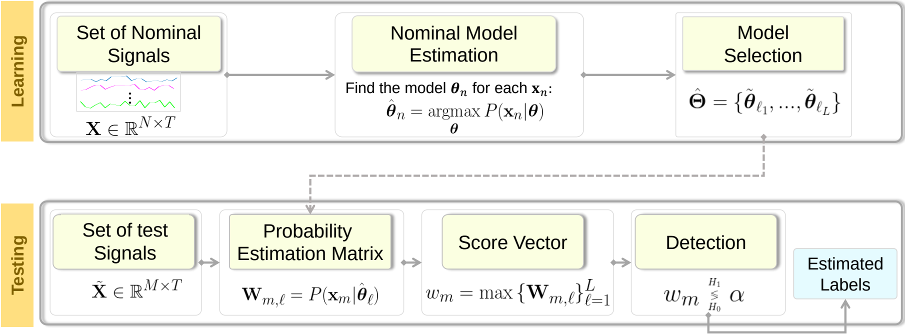
# HMM tasks

In general, there exists three main tasks related to the HMMs:

1. **Evaluation:** Given a HMM model $\theta$ and $x$, estimate the probability of observation: Find $P(x|\theta)$.
   Solution:
   - $P(x|\theta) = \sum_{all\ z} P(x, z|\theta)$. Forward-Backward Algorithm.
2. **Decoding:** Given a HMM model $\theta$ and observed sequence $x$, compute the hidden sequence that best models the observations: Find $z$.
   Solution:
   - $P(x, z|\theta) = P(x|z, \theta)P(z, \theta)$. Viterbi Algorithm.
3. **Learning:** Given the observed sequence $x$, estimate the most likely HMM model $\theta$ using the maximum likelihood method: Find $\theta$.
   Solution:
   - Choose $\theta = \{\pi, \mathbf{A}, \mathbf{B}\}$ such that $P(x|\theta)$ is maximized. Maximum likelihood estimation.

# HMM-Learning for Anomaly Detection

The proposed approach has two steps: Learning and Testing.




## Hypothesis

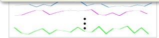
The following binary hypothesis test is considered to discriminate anomalies:

$$H_0 : \text{Absence of anomalies}$$

$$H_1 : \text{Presence of anomalies}$$

# HMM-Learning for Anomaly Detection

Set of Nominal Signals



$\mathbf{X} \in \mathbb{R}^{N \times T}$

Nominal Model Estimation

Find the model $\boldsymbol{\theta}_n$ for each $\mathbf{x}_n$:
$\hat{\boldsymbol{\theta}}_n = \underset{\boldsymbol{\theta}}{\text{argmax}}\, P(\mathbf{x}_n|\boldsymbol{\theta})$

Model Selection

$\hat{\boldsymbol{\Theta}} = \{\tilde{\boldsymbol{\theta}}_{\ell_1}, \ldots, \tilde{\boldsymbol{\theta}}_{\ell_L}\}$

Let $X = [x_1, \ldots, x_N]^T$ be a set of time series, where $x_n = [x_{n,1}, \ldots, x_{n,T}]$, with $x_n \in \mathbb{R}^T$.

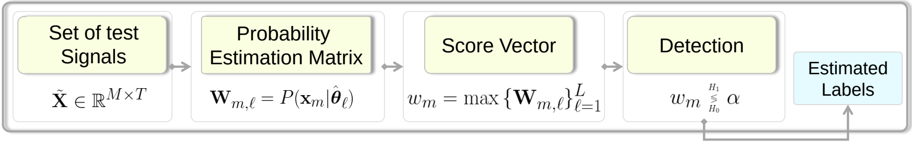
## Learning

- Learn the HMM model parameters $\boldsymbol{\theta}_n = \{\pi^{(n)}, \mathbf{A}^{(n)}, \mathbf{B}^{(n)}\}$ that maximize the log-probability of the observed sequence $x_n$:

$$\tilde{\boldsymbol{\theta}}_n = \arg\max_{\boldsymbol{\theta}}\; \log(P(x_n|\boldsymbol{\theta}_n)),\; n = 1, \ldots, N \tag{3}$$

  - $x_n$ : $n$-th time signal from the set $\mathbf{X} \in \mathbb{R}^{N \times T}$
  - Efficiently solved via the Baum-Welch algorithm
  - The set of learned HMM is written as $\boldsymbol{\Theta} = \{\tilde{\boldsymbol{\theta}}_1, \ldots, \tilde{\boldsymbol{\theta}}_N\}$
- Select $L$ HMM models for the testing as: $\{\tilde{\boldsymbol{\theta}}_{\ell_1}, \ldots, \tilde{\boldsymbol{\theta}}_{\ell_L}\}$, with $L \ll N$.

# HMM-Learning for Anomaly Detection




## Testing

- Estimate the probability of observation of $x_m$ by using $\{\tilde{\boldsymbol{\theta}}_1, ..., \tilde{\boldsymbol{\theta}}_L\}$:

$$\mathbf{W}_{m,\ell} = P(x_m | \hat{\boldsymbol{\theta}}_\ell), \tag{4}$$

  - $x_m$: $m$-th test signal, with $m = 1, ..., M$.
  - $\mathbf{W}_{m,\ell}$: matrix of probabilities of test sequences, with $\ell = 1, ..., L$, $\mathbf{W} \in \mathbb{R}^{M \times L}$.
- Compute the score vector $\mathbf{w} \in \mathbb{R}^M$ by selecting the maximum value of each row in $\mathbf{W}$.
- Discriminate potential anomalies in the score vector using the threshold $\alpha$.

Remark: Test signals with high probability, given the learned HMM model, are very likely to belong to the normal set (to be normal).

# Simulations and Results

## Evaluation Metrics

- Confusion matrix

| | | Prediction | |
|---|---|---|---|
| | | Normal | Abnormal |
| Groundtruth | Normal | True Negative (TN) | False Postive (FP) |
| | Abormal | False Negative (FN) | True Positive (TP) |

- Metrics for Correct Detection

$$\text{Precision} = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FP})}$$

$$\text{Recall} = \frac{\mathrm{TP}}{(\mathrm{TP}+\mathrm{FN})}$$

$$\text{F-score} = \frac{2\mathrm{TP}}{(2\mathrm{TP}+\mathrm{FP}+\mathrm{FN})}$$

*Metrics close to 1 means a good performance

## Methods to be compared

- OC-SVM [RBF Kernel]: One-Class Support Vector Machine
- HMAD: Hidden Markov Anomaly Detection.
- HMM-Learn: Presented here.

## Synthetic Dataset

A set of Gaussian sequences of $T = 600$ temporal points was generated. The set is divided as:

- Learning set: $N = 500$
- Testing set: $M = 100$

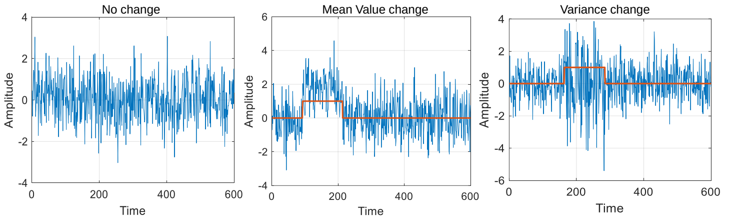
# Parameter setting

**Remark**: $X = [x_1, ..., x_n, ..., x_N]^T$ is the set of time series, and $x_n = [x_{n,1}, ..., x_{n,T}]$, with $x_n \in \mathbb{R}^T$.

**Scenarios:** For $u \neq 0$, $v \neq 1$, and $t_s$ : temporal segment.

| | Mean Value Changes | | Variance Changes | |
|---|---|---|---|---|
| | **Scenario 1** | | **Scenario 3** | |
| | Learning | Testing | Learning | Testing |
| $H_0$ | $x_n \sim \mathcal{N}(0,1)$ | | $x_n \sim \mathcal{N}(0,1)$ | |
| $H_1$ | | $x_{n,t_s} \sim \mathcal{N}(u,1)$ | | $x_{n,t_s} \sim \mathcal{N}(0,v)$ |

| | **Scenario 2** | | **Scenario 4** | |
|---|---|---|---|---|
| | Learning | Testing | Learning | Testing |
| $H_0$ | $x_{n,t_s} \sim \mathcal{N}(u,1)$ | | $x_{n,t_s} \sim \mathcal{N}(0,v)$ | |
| $H_1$ | | $x_n \sim \mathcal{N}(0,1)$ | | $x_n \sim \mathcal{N}(0,1)$ |



# Results from Synthetic Data

- Emission probabilities: Gaussian distributions (1).
- Introduced anomalies: $u = 1.6$ and $v = 1.9$.
- Fraction of anomalies: $10\%$ of the testing set.

Table 1: Summary of results from the synthetic dataset

| | Mean Value Changes | | | Variance Changes | | |
|---|---|---|---|---|---|---|
| | **Scenario 1** | | | **Scenario 3** | | |
| | Precision | Recall | F-score | Precision | Recall | F-score |
| OC-SVM | 0.684 | **1.000** | 0.813s | 0.400 | **1.000** | 0.571 |
| HMAD | 0.833 | 0.666 | 0.741 | 0.250 | 0.200 | 0.222 |
| HMM-Learn | **0.929** | **1.000** | **0.963** | **0.846** | **1.000** | **0.917** |
| | **Scenario 2** | | | **Scenario 4** | | |
| OC-SVM | NaN* | 0.000 | 0.000 | NaN | 0.000 | 0.000 |
| HMAD | **1.000** | **1.000** | **1.000** | 0.067 | 0.143 | 0.090 |
| HMM-Learn | **1.000** | **0.909** | **0.952** | **0.900** | **1.000** | **0.947** |

*NaN value: Anomalies no detected (TP = 0 and FP = 0).

# Case of study: Temporal Remote Sensing

**Dataset** [Provided by Florian Mouret]

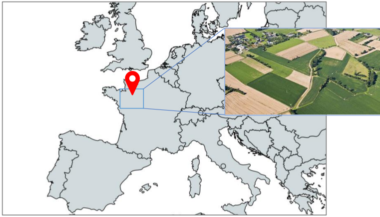
## Dataset Parameters:

- Area of study: Beauce, France
- Crops: Rapeseed
- Temporal resolution: 13 images from the Sentinel-2 satellite, corresponding to the harvest season of rapeseed (October 2017 to June 2018).

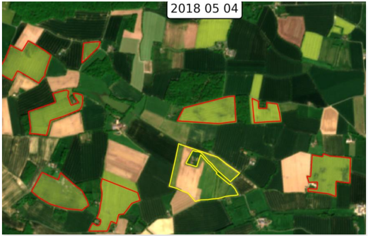
## Time Series Extraction

1. Parcel Extraction
2. NDVI estimation of each parcel
3. Median of NDVI

Total: $2218$ parcels (Time series)



Beauce, France




Rapeseed Parcels

# Anomalies in Rapeseed crops

## Anomalies in the Dataset

1. Late/Early growth
2. Heterogeneity
3. Late/Early senescence
4. Early flowering
5. Wrong shape

## Available labels (Total)

- Non-labeled: 1329
- Abnormal Labels (1): 725
- Normal Labels (0): 164

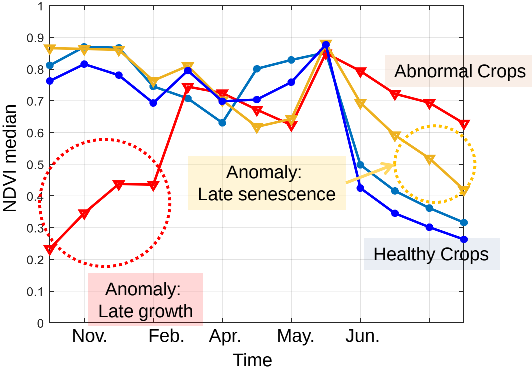
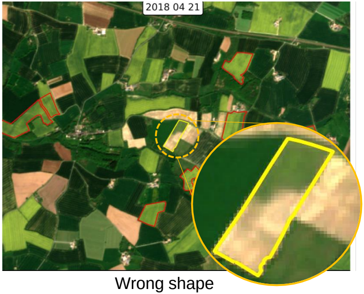
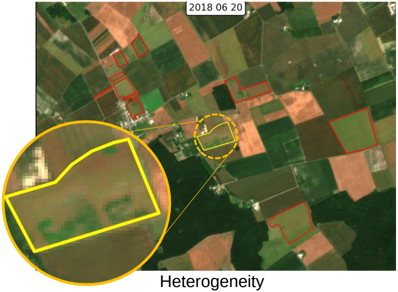



Illustration of normal and abnormal time series

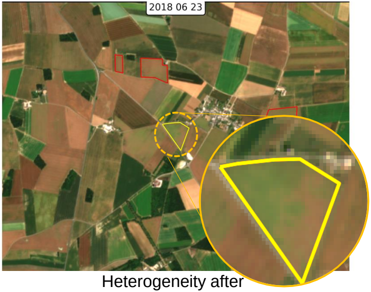
# Example of Anomalies in the Dataset




Wrong shape




Heterogeneity

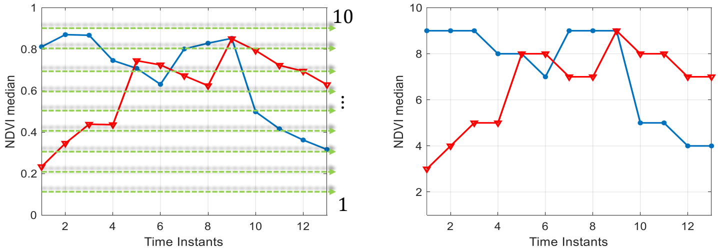



Heterogeneity after senescence

# Quantization

Given that the temporal resolution of the dataset is low ($T = 13$) Gaussian distributions are not suitable. Thus, the set of time series is discretized as follows:

- **Linear binning (regular):** The width of bins is 0.1. The codebook has 10 numbers, $K = 10$.

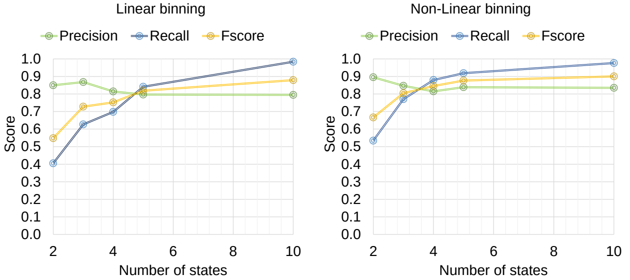



- **Non-linear binning:** Binning based on the median $\mu_G$ and standard deviation $\sigma_G$ of the *normal sequences* on each instant of time. The discretization can be written as

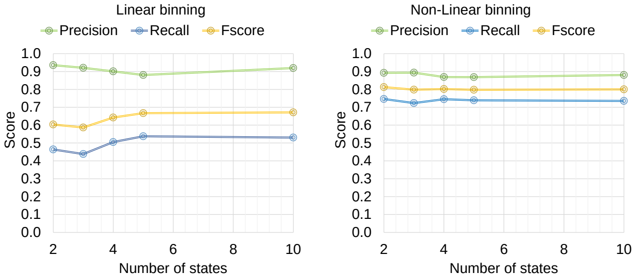
$$h(x) = \begin{cases} x_t = 1, & \text{if} \quad x_t < \mu_G^t - 3\sigma_G^t \\ x_t = 2, & \text{if} \quad \mu_G^t - 3\sigma_G^t \leq x_t \leq \mu_G^t + 3\sigma_G^t \\ x_t = 3, & \text{if} \quad x_t > \mu_G^t + 3\sigma_G^t \end{cases}$$

for $t = 1, ...., T$, where the number of elements in the codebook $K = 3$.

# Performance vs. Number of states




a) Single HMM model selection




b) Multiple HMM model selection

# Results from the Real Dataset

## Performance Comparison

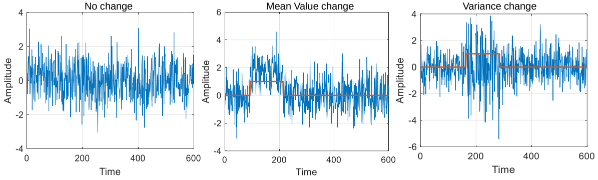
Table 2: Results from the different detectors based on the data with available labels.

| Detector | Precision | Recall | F-score |
|---|---|---|---|
| OC-SVM | 0.867 | 0.206 | 0.333 |
| HMAD | 0.873 | 0.492 | 0.629 |
| HMM-Learn | **0.894** | **0.724** | **0.799** |

# Conclusions (until now)

The work is not finished! However some conclusions are available:

- **A better performance in detection is obtained** in the anomaly detection approach based on HMM learning from time series.
- HMM learning for anomaly detection is **suitable for both high and low temporal resolution** sequences by properly selecting the distribution density.
- The proposed approach **allows the detection of anomalies related to mean value changes, variance changes, and even no changes**, as long as the learning step receives as input the signals considered as normal.



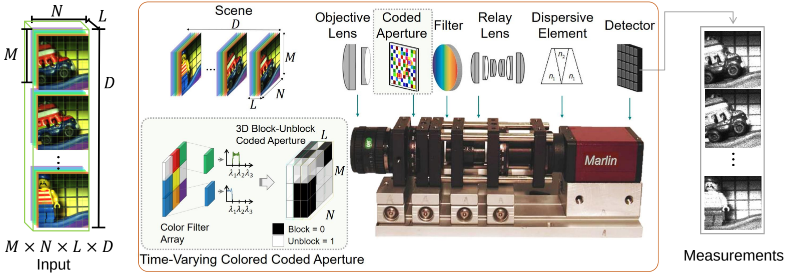
# Future Work

## Theory

- How to design the best HMM model selection in the current algorithm? (Minimum coherence, distance criterion).
- How to detect **when** occurs (time) the anomaly?
- To explore: Semi-Hidden Markov models (each state has variable duration), HMM states design.

## Applications

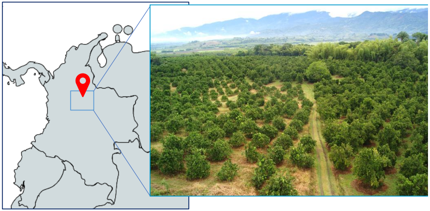
- Spectral video and compressive spectral video sensing



# Future Work

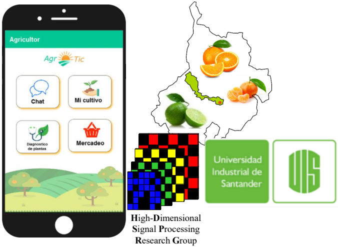
## Applications

- Apply the approach to citrus crops in Colombia: The main limitation is that there is few information about normal and abnormal citrus.



- Citrus project (ECOS-Nord 2019): Anomaly detection in citrus using Optical and SAR data.


- **AgroTIC**: App for the agriculture



# References I


**Anomaly Detection Surveys**

- V. Chandola, A. Banerjee and V. Kumar, Anomaly detection: a survey, *ACM Computing Surveys (CSUR)*, vol. 41, no. 3, pp. 1-62, 2009.
- M. A. F. Pimentel, D. A. Clifton and L. Tarassenko, A review of novelty detection, Signal Processing, vol. 99, pp. 215-249, 2014.

**Hiden Markov Models**

- L. R. Rabiner, A tutorial on hidden markov models and selected applications in speech recognition, Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, 1989
- B. Lorbeer, T. Deutsch, P. Ruppel, A. Küpper, Anomaly Detection with HMM gauge likelihood analysis, 2019 IEEE Fifth International Conference on Big Data Computing Service and Applications (BigDataService) , 2019.

**Algorithms used in Experiments**

- B. Schölkopf, J. Platt, J. Shawe-Taylor, A. Smola, and R. Williamson, Estimating the Support of a High-Dimensional Distribution, Neural Computation, vol. 13, no. 7, pp. 1443-1471, 2001
- N. Görnitz, M. Braun, and M. Kloft, Hidden Markov anomaly detection, in International conference on machine learning, 2015, pp. 1833-1842.

# References II

**AgroTIC Project**


- A. Camacho, and H. Arguello, Smartphone-based application for agricultural remote technical assistance and estimation of visible vegetation index to farmer in Colombia: AgroTIC. In Remote Sensing for Agriculture, Ecosystems, and Hydrology XX 2018 Oct 10 (Vol. 10783, p. 107830K). International Society for Optics and Photonics.

Thanks you for your attention!

**Questions?**

Merci!/Thanks!/¡Gracias!